Proyecto/Guía docente de la asignatura

Asignatura	ANÁLISIS DE DATOS CATEGÓRICOS		
Materia	ANÁLISIS DE DATOS CATEGÓRICOS		
Módulo	Básico		
Titulación	GRADO EN ESTADÍSTICA		
Plan		**Código**	47102
Periodo de impartición	Primer Semestre	**Tipo/Carácter**	OB
Nivel/Ciclo	GRADO	**Curso**	4º
Créditos ECTS	6.0		
Lengua en que se imparte	Castellano		
Profesor/es responsable/s	Agustín Mayo Iscar		
Datos de contacto (E-mail, teléfono…)	agustinm@eio.uva.es Tfno 983184170		
Departamento	Estadística e I.O.		

1. Situación / Sentido de la Asignatura

1.1 Contextualización

¿Qué es el ADC? "Análisis de Datos Categóricos" es ya un término acuñado dentro de la Estadística Aplicada que describe una gran cantidad de modelos estadísticos que explican estructuras de datos en los que las variables respuesta son discretas, ya sean estas numéricas, nominales u ordinales. ¿Por qué el ADC? Porque es preciso dar respuestas adecuadas, basadas en criterios científicos, a preguntas como las siguientes: ¿Cuál es la proporción de individuos de una población que padece SIDA?, ¿Es el AZT efectivo en el desarrollo de los síntomas de SIDA?, ¿Tiene la aspirina un efecto protector sobre el infarto de miocardio?, ¿Fumar produce cáncer de pulmón?, ¿Cuál es el grado de satisfacción de los consumidores de Mahou?, ¿Qué relación existe entre el nivel de ingresos y el nivel de estudios?, ¿Cuál es la relación entre el consumo de alcohol, cigarrillos y marijuana?, ¿Qué dosis de cypermetrina debemos aplicar para reducir a la tercera parte la población de heliotis virescens?, ¿Cambia el status ocupacional de padres a hijos?, ¿Por qué ocurrió la catástrofe de la nave espacial Columbia?, ¿Qué variables, y en qué medida, determinan la gravedad de un paciente ingresado en la UCI?, … El estudiante aprenderá la metodología estadística básica necesaria para dar respuesta a preguntas como las anteriores, y a otras muchas que de forma similar se plantean en todas las ramas de la actividad humana. La asignatura está orientada a las aplicaciones del ADC, y por ello una buena parte del trabajo que el estudiante tendrá que realizar será de índole práctico, mediante la utilización de herramientas informáticas y la interpretación de los resultados de los análisis que lleve a cabo, contribuyendo de ese modo a la adquisición del bagaje de "pensamiento estadístico" que todo profesional debe poseer.

1.2 Relación con otras materias

1.3 Prerrequisitos

Es recomendable conocer los elementos básicos de Probabilidad e Inferencia Estadística, así como de Álgebra y Cálculo Infinitesimal. Asimismo, es recomendable la capacidad de leer inglés técnico.

2. Competencias

2.1 Generales

Comprensión y capacidad para la puesta en práctica del complejo proceso del análisis estadístico de datos, desde la formulación del problema, el diseño y recogida de datos, hasta el ajuste, análisis y validación de modelos estadísticos, así como la interpretación de resultados y presentación de los mismos; todo ello en muy diversos contextos de aplicación, como pueden ser las ciencias sociales, la epidemiología, las ciencias de la salud o la industria. Esta capacitación general se alcanzará como resultado de las distintas actividades y de las aplicaciones que en campos diversos y de forma coordinada se llevarán a cabo en las asignaturas de este bloque.

Capacidad para la aplicación práctica de modelos paramétricos de regresión de índole muy diversa: modelos de regresión lineal, ANOVA y ANCOVA; modelos log-lineales, logísticos y de poisson, así como otros modelos lineales generalizados y no lineales; modelos de Cox, de tiempo de fallo acelerado y de riesgo aditivo en el contexto del análisis de supervivencia o de la fiabilidad; modelos lineales para respuesta multivariante.

Capacidad para hacer una valoración del ajuste y diagnóstico, así como de la comparación de modelos con los procedimientos adecuados.

Capacidad para interpretar los resultados del ajuste de cualquiera de los modelos en el contexto de cada problema real de aplicación, ya sea a través de las inferencias sobre coeficientes, sobre las predicciones o sobre otros parámetros de interés.

Capacidad para manejar las técnicas adecuadas para la resolución de problemas específicos en el ajuste de modelos como pueden ser las transformaciones de Box-Cox, la heterocedasticidad, la multicolinealidad o la sobredispersión.

Capacidad para la utilización de programas de estadística a un nivel avanzado y para el desarrollo de métodos no implementados en los programas estándar de estadística.

Capacidad para la resolución de problemas mediante técnicas de simulación y computación intensiva.

2.2 Específicas

Capacidad para el análisis de datos categóricos, ya sea la valoración del tipo de asociación en tablas de contingencia en diferentes condiciones, con el ajuste de modelos log-lineales, el ajuste de modelos logísticos u otros procedimientos específicos para respuesta catégorica, así como la interpretación de resultados.

3. Objetivos

Generales

Que el estudiante aprenda a reconocer problemas de respuesta discreta y a formular modelos estadísticos adecuados para su resolución.

Aprender el manejo de paquetes de programas estadísticos, como R o SAS, para el Análisis de Datos Categóricos.

Interpretar los resultados del ajuste de modelos para datos categóricos en problemas aplicados.

Aprender a seguir los diferentes pasos del proceso que va desde la formulación del problema real por otros profesionales, hasta la solución estadística y su comunicación.

Específicos

Que el estudiante aprenda a manejar los métodos estadísticos más usuales en tablas de contingencia 2x2, especialmente la comparación de proporciones, riesgo relativo, razón de ventajas, test exacto de Fisher, test de McNemar.

Conocer e interpretar los tipos de muestreo básicos asociados al estudio de tablas de contingencia, junto a las verosimilitudes asociadas y a los procedimientos de estimación y contraste subyacentes al ajuste de diferentes modelos.

Conocer, aplicar e interpretar el test CMH en el análisis de la independencia condicional en tablas 2x2xK, así como calcular los estimadores de la OR común bajo asociación homogénea.

Que el estudiante conozca la teoría básica del ajuste de modelos log-lineales en tablas de contingencia de diferentes dimensiones y sus aplicaciones al análisis de la asociación de variables categóricas.

Conocer los fundamentos del ajuste de modelos logísticos para una respuesta dicotómica cuando se tienen variables explicativas de diferente índole, interpretando los parámetros del modelo, estimando probabilidades y otras cantidades de interés como la ED50, la sensibilidad o la especificidad de una prueba diagnóstica.

Conocer, para una respuesta multinomial, la aplicación de modelos logit para respuesta nominal y de logits acumulativos para respuesta ordinal.

Conocer el uso de modelos de regresión de Poisson: la verosimilitud, la estimación de parámetros y su interpretación, estimación de medias y valoración del ajuste del modelo.

4. Contenidos y/o bloques temáticos

Bloque 1: "Nombre del Bloque"

Carga de trabajo en créditos ECTS:

a. Contextualización y justificación

La ya citada correspondiente a toda la asignatura

b. Objetivos de aprendizaje

Los ya citados correspondientes a toda la asignatura

c. Contenidos

1. Introducción a los problemas con respuesta categórica

i. Reconocimiento de problemas diversos cuya solución requiere del ADC, mediante la observación de diferentes ejemplos.

ii. Lectura y manejo de diferentes tipos de datos categóricos mediante R y SAS. Creación de tablas de frecuencias y porcentajes.

iii. El método de Wald para obtener intervalos de confianza y contrastar hipótesis, y su aplicación a la estimación de una probabilidad.

iv. Aplicación de métodos de estimación basados en el TRV (o Deviance) y en el Score a la estimación de una probabilidad. Test chi-cuadrado.

v. Problemas multiparamétricos.

2. Comparación de Proporciones y Tablas de Contingencia 2x2

i. Diferentes tipos de estudios. Estimación en estudios prospectivos y retrospectivos. Causalidad y asociación.

ii. Estimación de la diferencia de dos probabilidades (en muestras independientes) y del "Riesgo Relativo" (RR) utilizando distribuciones asintóticas.

iii. La "Odds Ratio" (OR) o razón de ventajas y su relación con el RR.

iv. Interpretación de la OR e inferencias asintóticas sobre la misma.

v. Utilidad de un diseño de muestras apareadas. Comparación de dos probabilidades (test de diferencia nula y estimación de la diferencia). Homogeneidad y Simetría en una tabla 2x2. Test de simetría de McNemar.

vi. Comparación de dos o más proporciones: Chi2 y TRV.

3. Tablas de Contingencia

i. Tablas de Contingencia. Muestreos de Poisson, Multinomial y Multinomial producto. Otros tipos de muestreo.

ii. La función de verosimilitud y la estimación máximo verosímil.

iii. Relaciones entre las distribuciones al condicionar por las marginales.

iv. Interpretación del modelo de no asociación en los tres tipos de muestreo básicos y su expresión formal. Presentación de otros modelos: cuasi independencia, simetría, homogeneidad marginal…, modelo saturado y modelo nulo.

v. Estimación máximo verosímil bajo no asociación.

vi. Tests de ajuste de un modelo: Test Chi2 y TRV (o Deviance). El AIC.

vii. Inferencias condicionales. Test exacto de Fisher.

viii. La paradoja de Simpson.

ix. Tablas 2x2xK. Asociación condicional y marginal. OR condicional. Asociación homogénea.

x. Test CMH de independencia condicional. Estimador MH de la asociación homogénea. Test de asociación homogénea.

4. Modelos Log-lineales en tablas IxJ

i. Introducción a los modelos log-lineales.

ii. Diferentes codificaciones y su interpretación.

iii. El modelo log-lineal como un modelo lineal generalizado.

iv. Inclusión de efectos dependiendo del tipo de muestreo.

v. Procedimientos para el ajuste de modelos log-lineales.

vi. Estimación de parámetros del modelo.

vii. Valoración del ajuste de modelos log-lineales. Cambio en la deviance para modelos anidados y el AIC.

viii. Ajuste de modelos adecuados a diferentes problemas: independencia, cuasiindependencia, simetría, cuasi-simetría, asociación uniforme, topológicos, efectos fila y/o columna,…

5. Modelos Log-lineales en tablas multidimensionales

i. Modelos log-lineales en tablas tridimensionales.

ii. Diferentes tipos de asociación en una tabla IxJxK: Independencia, independencia parcial, independencia condicional, asociación homogénea. Modelos log-lineales asociados.

iii. Estimación máximo verosímil.

iv. Inclusión de efectos de las marginales fijadas.

v. Ajuste de modelos log-lineales jerárquicos. Ruptura condicional de la deviance en modelos anidados.

vi. Alternativa al test CMH para contrastar la independencia condicional en tablas 2x2xK. Estimación de la OR común bajo asociación homogénea.

vii. Selección de un modelo log-lineal. Análisis secuencial de la deviance y eliminación de efectos. El AIC.

6. Modelos Logísticos

i. Problemas de respuesta binaria y predictores categóricos. Modelos logit y su relación con los modelos log-lineales.

ii. Ajuste de modelos logísticos.

iii. La tolerancia en problemas de respuesta-dosis: modelos logístico, probit y cloglog. Relación con los modelos lineales generalizados.

iv. Interpretación de los parámetros del modelo logístico. Interacciones.

v. Inferencias sobre los parámetros: EMV y su distribución asintótica, intervalos de confianza.

vi. Valoración del ajuste de modelos logísticos. Análisis de la deviance. El AIC. Análisis de residuos.

vii. Calibración (estimación de la dosis efectiva).

viii. Predicción. Reglas de clasificación (sensibilidad, especificidad,…curva ROC).

ix. Ajuste de modelos logísticos en estudios retrospectivos (caso-control).

x. Sobredispersión.

xi. Métodos exactos: inferencia condicional.

xii. Modelos para respuesta politómica.

7. Modelos de Poisson

i. Regresión de Poisson.

ii. Estimación de parámetros.

iii. Ajuste y selección de un modelo.

iv. Sobredispersión. Alternativa binomial negativa.

d. Métodos docentes

La asignatura se desarrollará en clases teóricas y clases prácticas. En estas últimas el estudiante dispondrá de ordenador y de materiales docentes que recrean situaciones reales y simuladas.

e. Plan de trabajo

f. Evaluación

Evaluación continua del trabajo realizado por el estudiante en las clases. Habrá un examen final de la asignatura.

g. Bibliografía básica

Agresti, A. (2013). *An Introduction to Categorical Data Analysis*. Third Edition. Wiley.
Collett, D. (2003). *Modelling Binary Data* (second edition). Chapman & Hall.
Simonoff, J. S. (2003). *Analyzing Categorical Data*. Springer-Verlag.

h. Bibliografía complementaria

Agresti, A (2002). *Categorical Data Analysis* (2nd. edition). Wiley.
Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley.
Le, C.T. (2010). *Applied Categorical Data Analysis and translational research* (2nd. edition). Wiley.

i. Recursos necesarios

j. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
6	Septiembre- Diciembre

5. Métodos docentes y principios metodológicos

6. Tabla de dedicación del estudiante a la asignatura

ACTIVIDADES PRESENCIALES	HORAS	ACTIVIDADES NO PRESENCIALES	HORAS
Clases teórico-prácticas	60		
Evaluación	2		
Total presencial	**62**		
Total presencial		**Total no presencial**	

7. Sistema y características de la evaluación

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Examen final de la asignatura	70%	
Evaluación del trabajo realizado por el estudiante en las clases y trabajos propuestos en clase..	30%	

CRITERIOS DE CALIFICACIÓN
• **Convocatoria ordinaria:** o … • **Convocatoria extraordinaria:** o **…**

8. Consideraciones finales